Noname manuscript No.
(will be inserted by the editor)

Viewing Computer Science through Citation Analysis

Salton and Bergmark Redux

Sitaram Devarakonda · Dmitriy Korobskiy · Tandy Warnow · George Chacko

Received: date / Accepted: date

Abstract w Computer science has experienced dramatic growth and diversification over the last twenty years. Towards a current understanding of the structure of this discipline, we analyze a large sample~~cohort~~ of the computer science literature from the DBLP database. For insight on the features of this cohort and the relationship within its components, we have constructed article level clusters based on either direct citations or co-citations, and reconciled them with~~to~~ major and minor subject categories in the All Science Journal Classification (ASJC). We describe complementary insights from clustering by direct citation and co-citation. ~~, and both point to the increase in computer science publications and their scope~~. Our analysis revealsshows cross-category clusters, some that interact with external fields, such as the biological sciences, while others remain inward looking.Overall, we document an increase in computer science publications and their scope.

[1] repositioned

Keywords Bibliometrics · Clustering · Research Evaluation · Computer Science · DBLP

Mathematics Subject Classification (2010) 01A85 · 01A90

Sitaram Devarakonda,
Netelabs, NET ESolutions Corporation, McLean, VA
E-mail: sitaramssd@gmail.com *Present address: Randstad USA, Atlanta, GA*

Dmitriy Korobskiy
Netelabs, NET ESolutions Corporation, McLean, VA
E-mail: dk@nete.com

Tandy Warnow
Dept of Computer Science, University of Illinois Urbana-Champaign, Champaign IL
E-mail: warnow@illinois.edu

George Chacko
Netelabs, NET ESolutions Corporation, McLean, VA
E-mail: netelabs@nete.com

1 Introduction

Computer science, and its applications, has experienced rapid growth and diversification over the last twenty years. As observed in a 2017 US National Academies Report, "A wide range of jobs in virtually all sectors demand computing skills to an unprecedented extent. And every academic discipline finds itself incorporating computing into its research and educational mission" [18]. More recently, the collective influence of the Internet of Things (IoT), 'big' data, accessible cloud computing, and advances in artificial intelligence have been presented as a driver for digital transformation~~postulated as a recent driver for growth and evolution~~ [24]. Given this rapid growth and expansion ~~powerful influence~~, an updated understanding of the present state and structure of computer science and its relationship to other fields can only inform planning and policy making at multiple levels from national level funding all the way down to faculty hiring strategy.~~could inform planning and policy making at multiple levels from national level funding all the way down to faculty hiring strategy.~~

In historical precedent, Salton and Bergmark conducted a study in 1979 of the computer science literature (419 computer articles published in 1974, and 3,812 references cited in these articles) [21]. Noting that that the scientific literature serves a rich source of information to study the structure and historical development of a field, these authors described the global structure of computer science as comprising three main areas: (i) theoretical foundations, such as theory of computation, (ii) hardware and computer systems, such as architecture, and (iii) software, such as programming systems. Related areas noted were (a) mathematics of computing, such as numerical analysis, (b) special software topics, such as operating systems, (c) data management and database systems, (d) methodologies valid for multiple applications, such as algebraic manipulation, (e) computer applications, such as computer graphics, and (f) non-technical aspects, such as computer education.

Looking beyond this historical triad of theoretical foundations, hardware and computer systems, and software, the Computing Classification System (CCS) published by the Association for Computing Machinery now consists of 13 top-level areas that reflect a more current view of the field [2]. This classification also addresses relationships with other fields under the category of Applied Computing. However, an easy way to map scientific publications to the CCS, especially interdisciplinary articles or those from proximal fields, does not seem to be available.

Other classification systems are available, such as the All Science Journal Classification (ASJC) developed and maintained by Scopus, the Web of Science research and categorical classifications from Clarivate Analytics, and the National Science Foundation classification system [19,11,9]. Scopus and the Web of Science also include comprehensive bibliographies with citation links and proprietary unique identifiers for publications. The Scopus All Science Journal Classification (ASJC), which we use in this study, is organized into 4 subject areas, 27 major subject areas, and 334 minor subject areas. All

three systems rely on applying one or more journal-derived labels to articles. A logical prediction, given the diversity of articles within journals, is limited specificity at the article level. Shu and colleagues recently noted in a comparative study of the Chinese Science Citation Database (CSCD) and the Web of Science that 46% of articles did not belong to the discipline of the journal they were published in [22]. Others have also discussed and critiqued disciplinary assignments using journal-based classifications [33,20]. Article classification systems have been constructed that escape some of the criticisms of journal-based classification [30,6,32], but do not seem to presently enjoy widespread use.

The purpose of this study is to develop an improved understanding of the structure of the field of computer science relative to the landmark study of Salton and Bergmark forty years ago. We extend their work by takimg advantage of modern bibliographic resources and clustering technologies and using a combination of article and journal approaches to study trends in the computer science literature. We also consider connections to other fields, especially biology, since approximately 42% of Scopus is classified under the top level subject areas of Life Sciences and Health Sciences. As a source of computer literature, we use DBLP, a reference bibliography for computer science [29]. The DBLP bibliography covers publications from computer science and includes publications from hybrid fields, where they are considered pertinent to computer science research. ~~In this study, we extend the work of Salton and Bergmark through a combination of article and journal approaches to study trends in computer science, while also considering connections to other fields, especially biology since approximately 42% of Scopus is classified under the top level subject areas of Life Sciences and Health Sciences. As a source of computer literature, we used DBLP, a reference bibliography for computer science [29]. The DBLP bibliography covers publications from computer science and includes publications from hybrid fields, where they are considered pertinent to computer science research. We describe below the high-level interactions of computer science within the Physical Sciences, and with the Social Sciences, Life Sciences, and Health Sciences.~~

We XXX XXXX XXXXX

2 Materials and Methods

Overview The underlying assumption is that, notwithstanding incomplete coverage, records in DBLP are greatly enriched for computer science [29]. Therefore, the contents of DBLP are an excellent sample of the computer science literature to analyze. The workflow described in the following paragraphs of Materials and Methods, consists at a high-level of (i) merging the DBLP and Scopus datasets using digital object identifiers(DOIs) (ii) extracting cited references for the articles common to DBLP and Scopus (iii) clustering by direct citation and by our modification of co-citation based clustering. The outcome was a dataset of 8,000,411 publications.

[gc 1] Reviewer 1

Our working definition of the computer science literature, for the purpose of this study, was all publications in the DBLP bibliography that (i) had a digital object identifier DOI, and (ii) could also be matched to article identifiers in the Scopus bibliography.~~For the purpose of this study, our working definition of the computer science literature was all publications in the DBLP bibliography that (i) had a digital object identifier (DOI), and (ii) could also be matched to article identifiers in the Scopus bibliography.~~ Cross-matching DBLP publications to records in the Scopus abstract and citation database of peer-reviewed literature enables us to harvest the richer links in Scopus, as well as to extract links to publications from other disciplines. Cross-matching to Scopus also allows the use of journal-based classifications when clustering documents at the article-level. We used DBLP articles from journals and conference proceedings to construct article clusters by using either direct citations or co-citations as links. We reconciled these clusters with the All Science Journal Classification (ASJC) developed and maintained by Scopus through a combination of automated and manual procedures, producing a dataset of 2,685,356 publications, which when combined with cited references extracted from Scopus grew to 8,000,411 publications.

*DBLP data*A stable release of the DBLP computer science bibliography [29] consisting of 7,079,994 records was downloaded as dblp-2018-08-01.xml.gz. Slightly over 95% of the publications within were published after 1996. Publications were parsed from the XML source file and loaded into a PostgreSQL database.

Scopus data As part of implementing a larger data platform for research evaluation [16], we have previously parsed the Scopus dataset, presently at over 88 million publications, into a custom schema in a PostgreSQL database. The total number of publications in Scopus labeled with major subject area Computer Science (in turn a subset of the Physical Sciences subject area) is 5,835,160.

Merging and graph construction Records in the DBLP dataset were matched to Scopus identifiers using digital object identifiers (DOIs). This procedure resulted in a dataset of 2,685,356 DBLP publications with Scopus identifiers where 1,278,322 (47.6%) were labeled as article and 1,407,034 (52.4%) as conference proceedings. References cited by these publications were then extracted from Scopus (7,129,006 records), resulting in a total of 8,000,411 publications and references.

We represented these 8,000,411 records, referred to as the *comp* dataset (Fig. 1, Table 1), as a graph where the 8,000,411 nodes represent publications and references and the 44,296,381 undirected edges represent citations within the dataset.

Clustering Clustering of publications is commonly accomplished through direct citation, bibliographic coupling, and co-citation, with direct citation being proposed as as the best approach to concentrate citation links [14,15]. Accordingly, we used direct citation links as the basis for cluster formation, and also co-citation to obtain an alternative view. In applying both clustering by direct citation and by co-citation, we attempted to consider, wherever possible,

the criteria articulated by Šubelj, van Eck, and Waltman [31] that (i) the largest cluster should be no more than 10 times the smallest one, (ii) small clusters should be eliminated, (iii) small changes and replicates should yield similar results ("stability"), (iv) computing time should be minimized where possible, and (v) the clustering should seem reasonable on a qualitative level ("intuitive sensibility").

Direct Citation Graclus [10] is a spectral graph clustering package that optimizes various clustering criteria, including normalized cut, ratio cut, and ratio association, and that has previously been applied to citation data [31]. We used v1.2 in our experiments. The *comp* dataset was formatted as an undirected graph, stored in a file with a header line indicating the number of nodes and edges, and used as input to Graclus, which requires the number of clusters to be formed as an input parameter. In preliminary experiments, we varied the number of clusters to be formed between 10 and 50 clusters (data not shown). At around 20 clusters, clusters size was relatively stable with the largest cluster containing roughly 10 times the number of nodes in the smallest one, so that 20 clusters is a good choice with respect to the criteria specified in [31]. Consequently, we used Graclus to generate 20 clusters, labeled 0-19 (Table 2), analogous to Level 1 of Waltman and van Eck's mapping of nearly 10 million publications but focused on the DBLP bibliography rather than a broader Web of Science sample [32].

We also used conductance, as defined in Shun et al. [23], to evaluate clustering by direct citation (smaller is better), noting that conductance has been found to be a good metric for this purpose [12,1]. In our analysis, we saw that the last cluster (cluster 19) had a much larger conductance value than the other clusters and also had the smallest number of nodes. We then examined results obtained using Graclus with two other numbers of clusters (18 and 22), and in each case, the highest numbered cluster had the greatest conductance value and also the smallest number of nodes. These results suggest that Graclus produces a final cluster that effectively serves as a container for 'left over publications' during the clustering procedure. Therefore, we limited our consideration of cluster 19 (the final of the twenty clusters) when interpreting results. The remaining 19 clusters had conductance values ranging from 0.09 to 0.25 with a median conductance of 0.15 (Fig. 2, Table 1).

Co-Citation. For an alternate view of these DBLP data, we constructed clusters using co-citation, the frequency with which a pair of articles is cited by other articles [26,17]. Co-citation, first described independently by Small and Marshakova in 1973 [17], provides insight into the emergence of new ideas derived from the association of previously independent ones. Unlike clustering by direct citation, where every input publication is assigned to a cluster and every citation is weighted equally, the co-citation relationship between papers is weighted to represent the strength of the co-citation history. Because this produces a weighted graph, clustering methods that address weights are required. Clustering by co-citation also considers weak inter-cluster interactions that involve modifications to standard clustering approaches. [3,7,27,28].

We used a modification of variable level clustering combined with agglomerative clustering, an approach developed in 1985 by Small and Sweeney [28] for co-citation analysis. Variable level clustering involves applying a threshold (below which all edges in a graph are deleted) then iteratively selecting edges with the highest normalized co-citation value and extracting connected components from the graph as clusters for each edge in turn. Three parameters are needed: (i) a threshold or starting level based on a quantile of normalized co-citation frequency, (ii) a level increment, and (iii) a maximum cluster size. An issue is the generation of very large clusters by chaining via low edge weights. Thus, at each iteration, any cluster exceeding the maximum cluster size is returned to the process and a higher threshold is applied to break such clusters.

We began by identifying highly cited articles in *comp* and selected those in the 90th percentile (212,311 articles). We then identified 4.3 million publications in Scopus that cite these 212,311 articles. For each of the 4.3 million citing publications in turn, all possible $\binom{n}{2}$ reference pairs were generated from a publication's cited references, where n is the number of references in a publication. The cited reference pairs this generated were then restricted to those where both members of a pair were in the set of 212,311 highly cited papers previously identified. A total of 46,463,117 unique co-cited pairs were thus obtained. The frequency of these co-cited pairs was then computed across the *comp* dataset and normalized using Salton's cosine formula [21] to limit dominance by areas with high citation activity. These data were represented in a graph where each node was a publication and the weighted edge between the pair was the normalized co-citation frequency.

In our implementation of variable level clustering (Fig. 3(a)), we set initial parameter values as follows: the threshold t is initially set to the median normalized co-citation frequency (quantile=0.5), increment $i = 0.1$, and maximum cluster size, $mcs = 200$. Thus, at the start, all edges below the median normalized co-citation frequency were deleted. Clusters were formed by assembling connected components from each co-cited pair beginning with the heaviest edge weight. Clusters below size 100 were retained and any cluster larger than 200 nodes was carried over to the next round. The threshold, t, was then incremented by 0.1 and the process repeated while progressively incrementing t. We used a bi-phasic approach where in which t ranged from 0.5–0.9, after which i was reduced to 0.01 for the range $0.9 \leq t \leq 0.99$. A final threshold of t=0.999 was applied to break the single remaining large cluster. Using this approach, 22,232 clusters containing 84,591 nodes were generated, each containing less than 100 nodes. Clusters containing only two nodes were discarded, bringing the total number of clusters down to 10,298. The publications in these 10,298 clusters were overwhelmingly drawn from the Physical Sciences (one of the four top level categories in the Scopus ASJC classification), of which computer science is a sub-category (Fig. 3).

Agglomerative clustering was then performed on these 10,298 clusters to generate higher-order clusters. To focus on larger clusters, only those with at least 10 nodes were used as input. Briefly, each cluster was now treated as a

node and the edge weight between two clusters was assigned to the maximum edge weight of all edges between the nodes in the two clusters. Edges were arranged in descending order. The first pair of clusters was merged and its edge weight with other interacting clusters was recalculated, again based on maximum edge weight. All edges were then re-ordered as before and the next pair of clusters was merged. The process was halted after 600 rounds to prevent large outlier clusters being generated (Fig. 3(d)).

3 Results

In our high-level study of the structure of a twenty year sample of the computer science literature, we chose to use both traditional journal-based and article-based approaches. Considering that traditional disciplines 'may only partly reflect the actual organization of today's scientific research' [32], we constructed article clusters at high levels of aggregation using citations to examine the computer science literature and also mapped these clusters to journal-based categories to take advantage of both article level and journal level approaches.

Data~~The process of selection and matching resulted in a dataset of publications (Materials and Methods). Of 4,291,130 DBLP publications with DOIs, only 2,685,356 had corresponding DOIs in Scopus.~~ Of 2.68 million publications in *comp*, approximately 2.07 million were assigned ASJC codes in Scopus corresponding to Computer Science (major subject area with 13 minor subject areas (Table 1), with the balance of 610,000 publications non-exclusively shared between 26 different major subject areas ranging from Engineering (330,048) to Dentistry (Fig. 1). The set of 2.07 million publications classified under the major subject area Computer Science in Scopus spanned all 13 minor subject areas with Software at 30.3% being the largest component and Computer Science (miscellaneous) at 0.9% the smallest. Publications in the *comp* dataset labeled Theoretical Computer Science (409,082) are classified under the the major subject area of Mathematics rather than Computer Science (Table 1).

Clustering by Direct Citation We constructed article-level clusters of this computer science dataset at a sufficiently high level of aggregation to avoid cognitive challenge and cross-matched them to the Scopus ASJC classification. To focus on relatively high signal, we only considered Scopus ASJC minor subject area categories that accounted for at least 15% of the publications in each cluster.

Figure 4 permits examination of these data from two perspectives: (i) rows: the clusters that map to a given ASJC minor subject area and (ii) columns: ASJC minor subject areas that comprised at least 15% of the publication in a cluster. Under these conditions of clustering and this threshold of 15%, 31 of the 334 ASJC minor subject areas are represented. Unsurprisingly, the broad categories Computer Science Applications, Software, and Electrical and Electronic Engineering register in 16, 12, and 10 clusters respectively, while Artificial Intelligence mapped to 7 different clusters. At the other end of the

range, 12 of the 31 ASJC minor subject areas were each detected only in a single cluster.

From the alternate perspective (columns), Cluster 17 was the most diverse and contained publications annotated with 8 minor subject area labels: Biochemistry, Chemistry(all), Genetics, Molecular Biology, Statistics & Probability, Computational Theory & Mathematics, and Computational Mathematics. Cluster 19 mapped to two areas but was excluded from qualitative analysis because of its high conductance value. Of the remaining clusters, Cluster 2 represents interactions between the four minor subject areas Theoretical Computer Science, Discrete Mathematics and Combinatorics, Applied Mathematics, and the more generic Computer Science (all). Clusters 3 and 4 include Computer Networks and Communications, and Cluster 18 (Artificial Intelligence, Cognitive Neuroscience, and Neurology) and clusters 5–9 include Management Science, Operations Research, Information Systems, Modeling and Simulation, and Human-Computer Interaction.

These data suggest that fields central to computer science in 2019 (Salton and Bergmark's historical triad of hardware, software, and theory) are more likely to be found in multiple clusters than peripheral fields. A second inference is that, in some cases, journal-based classification and our article clusters align fairly well (Hardware and Architecture). A third inference is that the ASJC minor subject area "Computer Applications" is relatively broad, and publications thus labeled are present at the >= 15% level in 16 out of 20 clusters. Finally, for this DBLP dataset, as we clustered it, interactions with fields outside computer science such as Biology (i.e., Biochemistry, Neurology) are detected in two separate clusters. The first appears to be the interaction of Biochemistry, Molecular Biology, and Genetics with Statistics, Mathematics, and Computer Science, and the second is the interaction of Neurology, Cognitive Neuroscience, and Artificial Intelligence.

These clusters cannot be easily characterized by using the CCS classification. For example, we manually matched the top 25–50 most heavily cited publications in each cluster to corresponding categories in the CCS. This was feasible with clusters 0 and 1 mapped reasonably well to the top level categories Hardware and Computer Systems Organization, but in other cases, the top cited papers often derived from biology, yet biology was clearly not representative of the majority of the nodes in these clusters.

Clustering by Co-citation. For an alternate examination of the data, we used co-citation frequencies to cluster the DBLP dataset as described above (Materials and Methods). Figure 5 shows a heatmap in which clusters constructed by co-citation are mapped to Scopus ASJC minor subject area labels, with the top subfigure showing results for those labels that account for at least 15% of the publications in a cluster, and the bottom subfigure showing results for those labels that account for at least 10% of the publications.

At the threshold of 15%, only 17 of 20 co-citation clusters mapped to at least one minor subject areas. At the 15% threshold, no co-citation cluster maps to more than three minor subject areas, in contrast to a maximum of eight minor subject areas for clustering by direct citation (Fig. 3, Cluster 17).

The threshold had to be reduced to 10% for all 20 co-citation clusters to map to at least one minor subject area. We interpret these results as indicative of broader clusters created by weaker linkages that accumulate during the agglomerative clustering phase [28].

Clustering by co-citation begins, for each cluster, with a pair of nodes that nucleates its subsequent formation. Thus, we also designated the pair of nodes in a cluster with the strongest edge as its nucleating pair and labeled the cluster by manually labeling the nucleating pair of documents. These labels show a high degree of correspondence to the ASJC minor subject area that accounts for the largest fraction of the nodes in a co-citation cluster (Table 4). Thus, using the nucleating pair as the basis for labeling co-citation clusters (as we generated them) may be valid, although its usefulness is likely to vary according to the data being examined; we also note that manual annotation (although beneficial) is not scalable. We provide the DOIs of these nucleating pairs matched to manually assigned labels for independent review (Table 3).

To examine the correspondence between clusters generated by direct citation vs co-citation, we mapped the contents of these clusters to each other (Fig. 6). At a threshold of 15%, similar to other cross-matching, 90% (18/20) of the co-cited clusters mapped to 1 or 2 direct citation clusters. This suggests that the majority of co-cited pairs tend to lie within the same cluster of publications linked by citations and, by extension, tend towards disciplinarity rather than interdisciplinarity.

4 Discussion

Considering expansion and diversification of the field of computer science, we revisited its characterization by Salton and Bergmark in 1979 [21]. In comparison to [21], we analyzed considerably more data, 2.68 million publications versus 391 by using two bibliographic databases, DBLP and Scopus, consisting of approximately 7 million and 88 million publications respectively. By linking the two datasets, we were able to harvest citation data as well as other metadata that enabled theo construct article level clusters in two different ways and reconciliation with the Scopus ASJC classification, for a journal level perspective.

Reconciling direct citation clusters to the Scopus ASJC classification yielded partially overlapping results that are consistent with the observation of Waltman and van Eck (2012) [32] 'that traditional disciplines such as those just mentioned only partly reflect the actual organization of today's scientific research'. Of interest to us was the single obviously multidisciplinary cluster in which at least 15% of its component publications were labeled with the ASJC minor subject areas Biochemistry, Chemistry, Computational Mathematics, Computational Theory and Mathematics, Computer Science Applications, Genetics, Molecular Biology, Statistics & Probability. A second cluster mapped to Artificial Intelligence, Cognitive Neuroscience, and Neurology. Both suggest collaboration between computer science and biology. At the opposite end of

this spectrum is the cluster that maps to Hardware and Architecture, Electrical and Electronic Engineering, and Software. These data suggest that, at least from the perspective of a high level of aggregation, some subfields within computer science may be primarily inward looking (remain concerned with fundamental questions in computer science and electrical engineering), while others are more actively engaged with fields external to computer science.

A comparison of the clusters generated using direct citation and co-citation shows interesting contrasts. The nodes in a co-citation cluster often map largely to one or two direct citation clusters: for example, 98% of the nodes in co-citation cluster 19008 map to direct citation cluster 8, which in turn aligns with theory, software, applications, and networks. Conversely, co-citation cluster 18947 nucleated by a pair of articles in the Journal of Applied Mathematics and Computation is distributed between direct citation clusters 11, 16, and 17, effectively spanning Artificial Intelligence, Applied Mathematics, Biochemistry, Chemistry, Genetics, Theory, Software, Applications, and Statistics.

The focus of this article was on high-level features, and a preference for simplicity and intuitiveness in the choice of methods. This preference may address questions regarding generalizability and the specific choices we made, such as the (i) use of the DBLP dataset matched to Scopus, which may not capture all aspects of computer science and its interactions with other fields, (ii) the basis for clustering, and (iii) mapping the results of this clustering against a classification designed around journals rather than individual articles. We believe that an approach that combines journal-level with article-level analyses is useful for studies of this kind. However, we acknowledge that our approach would not detect emerging fields with few publications that could be important signs of innovation.

We speculate that biology, often dominant in bibliometric studies, is restricted to two clusters on account of (i) the focus of DBLP, (ii) our use of normalized co-citations, and (iii) the threshold set for detection. Future research should, of course, include complementary investigations at finer levels of granularity and sensitivity using article-level and topic approaches that others have developed [13,5,4,6,25,30]. We also refer readers to a related study of the computer science literature [8] that is focused on evolving interdisciplinarity in computer science using data from Microsoft Academic Research and a classification of Computer Science into 24 categories.

In the 40 years since Salton and Bergmark's landmark paper, the field of computer science has not only expanded in volume, it has expanded in its interactions with other fields, and has also resulted in new disciplinary and interdisciplinary subfields. Furthermore, machine learning and data science, which build off computer science and statistics, are emerging as major fields that are driving innovation in industry and science, and research in these areas is increasingly being performed in fields external to computer science.While DBLP provides an insight into what is commonly accepted as computer science, additional evaluation of the broader literature that uses and develops computer science is needed to better assess the impact of computer science.

5 Figures

Figures

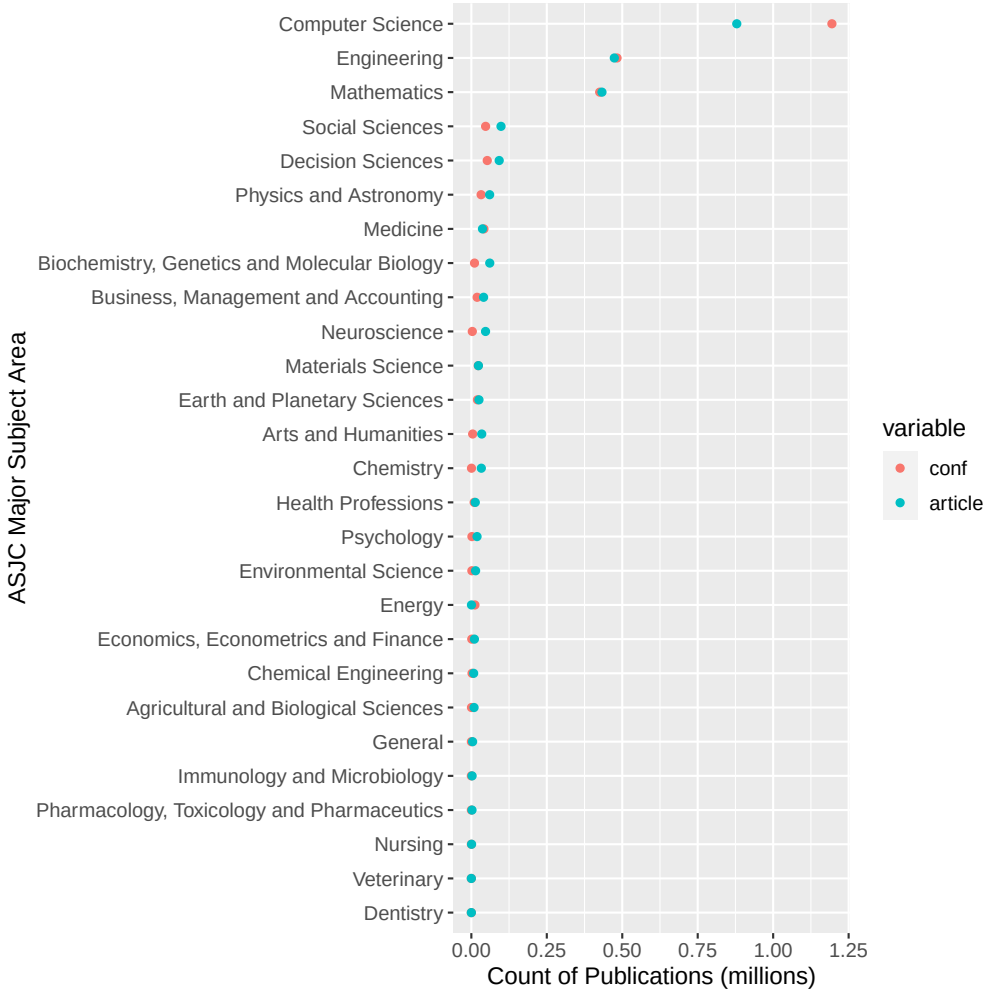

Fig. 1 Summary of DBLP data cross-matched with Scopus. 2,685,356 publications from DBLP were cross-matched with Scopus and then grouped by the 27 major subject areas in the ASJC (Scopus) classification. The largest number of publications are contributed by Computer Science; Engineering; Mathematics; and then by Social Sciences; Decision Sciences; Physics and Astronomy; Medicine; and Biochemistry, Genetics, and Molecular Biology. Publications were further annotated with respect to being either articles *(ar)* or conference proceedings *(cp)*. For this dataset, the major subject area of Computer Science with 1,194,623(cp) & 879,396 contributed the most publications while Dentistry with 0 (cp) & 1 (ar)) contributed the least.

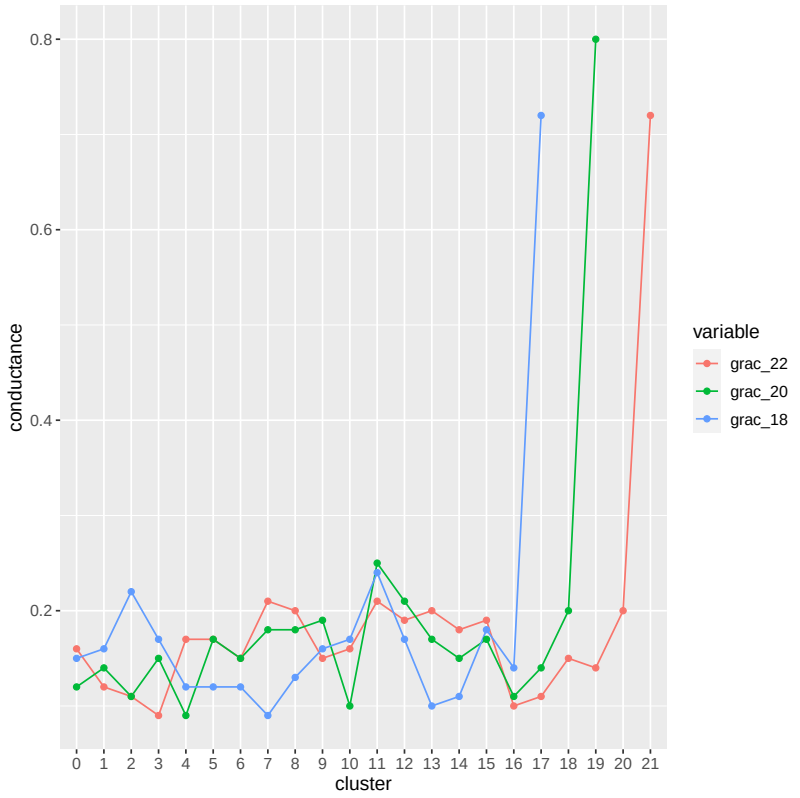

Fig. 2 Conductance Measurements of Clusters Generated by Graclus of the direct citation dataset. 2,685,356 DBLP publications, 7,129,006 cited references, and 44,296,381 citations were clustered using Graclus into 18 (grac_18), 20 (grac_20), or 22 (grac-22) clusters. Conductance, $\phi(S)$, was measured for these clusters considering only the edges between publications using the formula: $\phi(S) = |\partial(S)|/min(vol(S), 2m - vol(S)$, where $\partial(S)$ is the boundary (number of edges leaving a set), vol(S) is volume of a set of vertices as the sum of the degrees of the vertices in a set, and m is the number of undirected edges in a set [23].

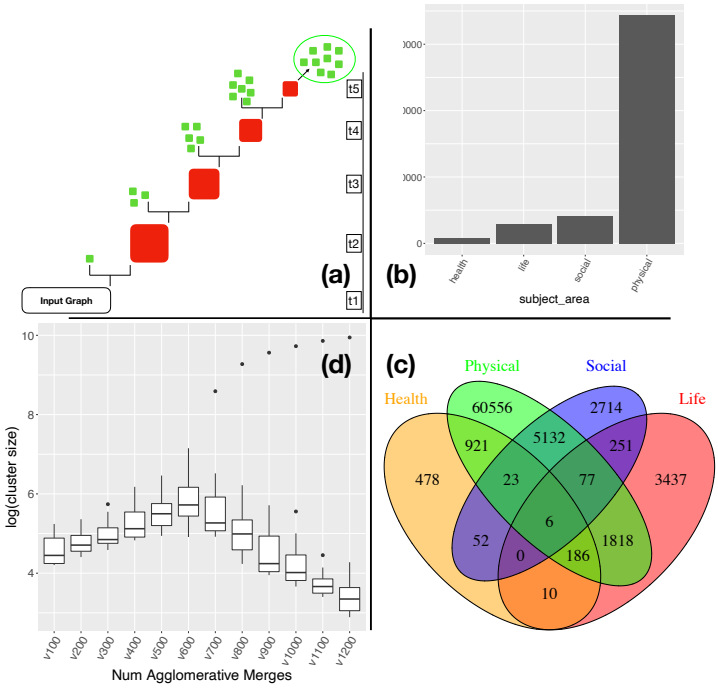

Fig. 3 Co-citation analysis. (a) Schematic representation of variable clustering protocol modified from Small and Sweeney (1985) [28]. Three parameters are specified: (i) a threshold or starting level based on a quantile of normalized co-citation frequency, (ii) a level increment, and (iii) a maximum cluster size. Input data is a set of co-cited publications with edge-weight defined by normalized co-citation frequencies. Green clusters are within the max cluster size. At the initial threshold, $t1$, a single cluster below the maximum cluster size, *mcs* (green), along with one large cluster above it (red) are generated. As the threshold is incremented to $t2$, additional clusters of acceptable size is generated. The cascade continues to completion, which is defined by all clusters being of size less than or equal to the *mcs*. In this schematic, five rounds are adequate for the process to run to completion. (b) The distribution of publications (using fractional counting) across four top-level ASJC subject areas after applying variable level clustering as in a) (c) The Venn diagram of the fractional counting given in (b). (d) The distribution of cluster sizes (logarithmic y-axis) as a function of the number of iterations of the agglomerative clustering technique; note that the largest cluster is extremely large when the number of iterations exceeds 600.

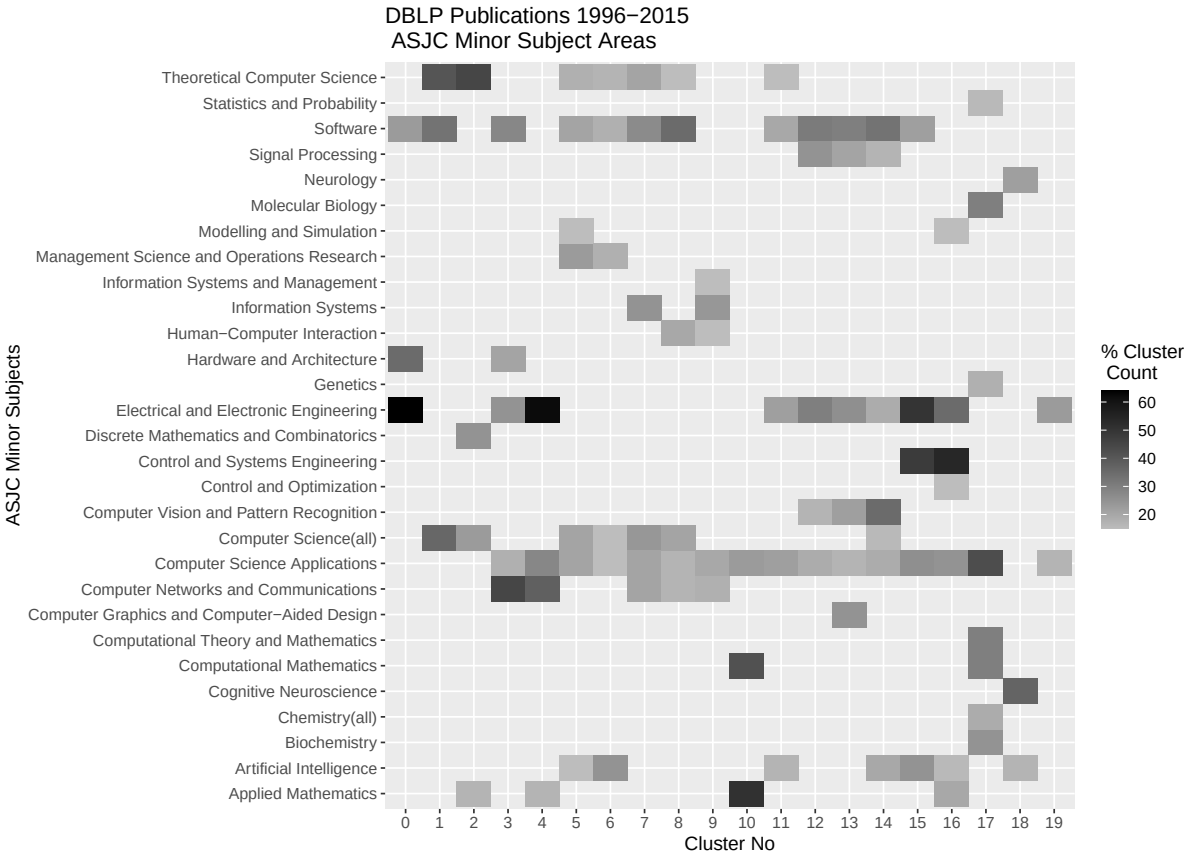

Fig. 4 Heat map for the clustering obtained by direct citation. The y-axis (rows) correspond to topics, defined by Scopus characterizations, and the x-axis (columns) represent the 20 different clusters. Each cluster is characterized by topics that label at least 15% of the publications in the cluster.

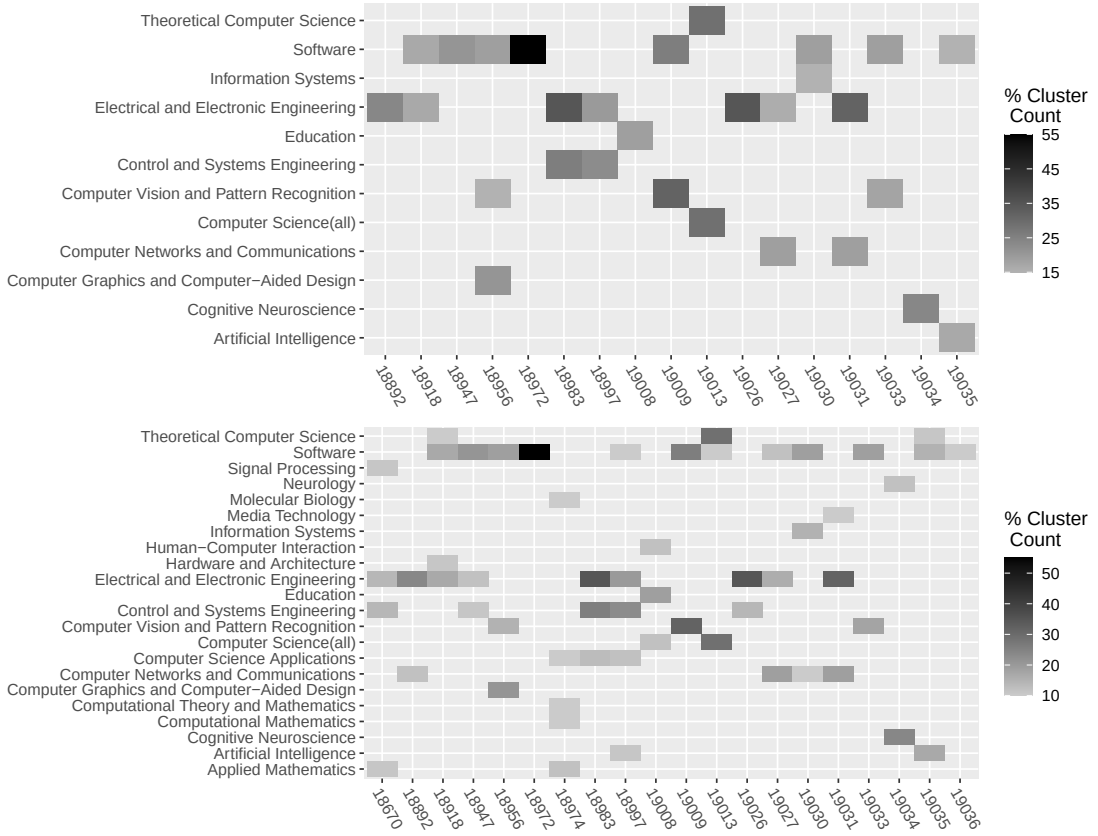

Fig. 5 Heat map for the clustering obtained by co-citation, using two thresholds for inclusion (top: 15%, bottom: 10%). The y-axis (rows) correspond to topics, defined by Scopus characterizations, and the x-axis (columns) represent the 20 different clusters. Each cluster is characterized by topics that label at least the required minimum percentage of publications in the cluster (top: 15%, bottom: 10%).

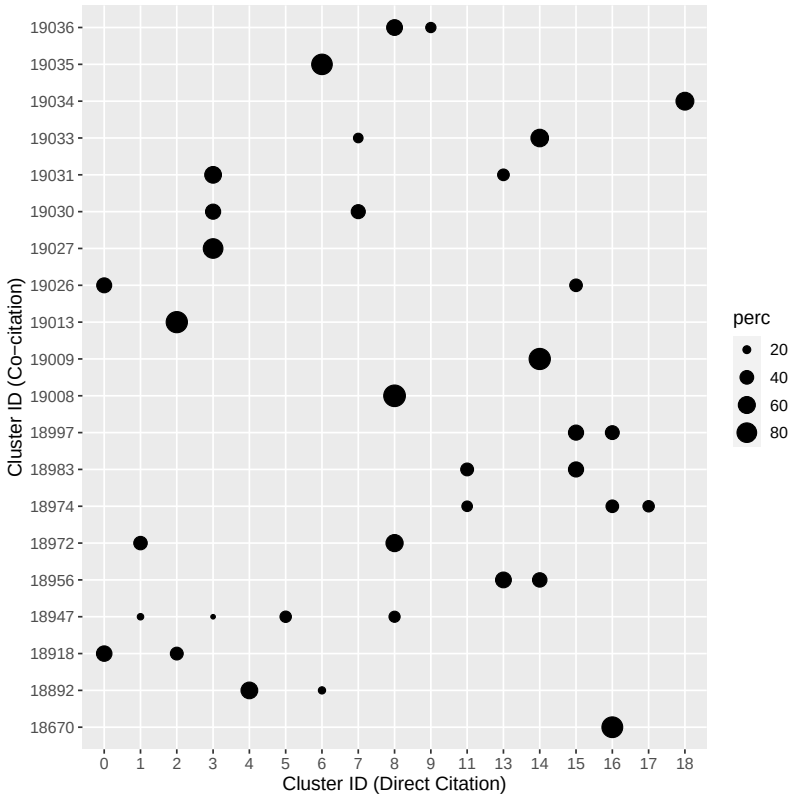

Fig. 6 Intersections between clusters generated using direct citation and co-citation features. x-axis: Clusters generated by Graclus number 0-19. y-axis: Cluster generated by variable level clustering and agglomerative clustering using a modification of Small and Sweeney (1985) [28]. Point size (perc) is the percentage of a co-cited cluster that maps to a corresponding Graclus cluster. A minimum threshold of 15% was set. Graclus cluster 19, the 20th cluster, did not map to any cluster on the y:axis.

Tables

Table 1 Characterization of the *comp* dataset relative to Scopus ASJC minor subject areas.

minor_subject_area	Percent of Publications
Software	30.1
Computer Science Applications	25.0
Computer Networks and Communications	21.5
Theoretical Computer Science	19.6
Computer Science(all)	19.4
Artificial Intelligence	14.2
Information Systems	11.8
Hardware and Architecture	11.8
Signal Processing	9.8
Computer Vision and Pattern Recognition	9.4
Computational Theory and Mathematics	7.9
Human-Computer Interaction	7.4
Computer Graphics and Computer-Aided Design	5.2
Computer Science (miscellaneous)	0.9

Table 2 Descriptive statistics of the 20 clusters produced from the *dataset* using Graclus (direct citation). The number of publications in each cluster, the conductance [23] of each cluster, the total number ASJC minor subject area labels assigned to publications in each cluster and the number of unique labels in each cluster are shown.

Cluster	Publications	Conductance	Total ASJC Labels	Unique ASJC Labels
0	111,294	0.12	265,664	142
1	117,057	0.14	246,960	166
2	116,251	0.11	280,602	165
3	353,366	0.15	881,693	200
4	145,081	0.09	349,020	154
5	92,097	0.17	248,168	186
6	71,865	0.15	199,681	163
7	179,927	0.18	465,181	202
8	302,656	0.18	760,117	214
9	69,520	0.19	197,031	174
10	42,462	0.10	102,838	141
11	448,030	0.25	1,229,061	224
12	70,738	0.21	216,546	179
13	105,232	0.17	289,318	187
14	199,176	0.15	551,657	208
15	64,384	0.17	195,679	167
16	89,340	0.11	240,157	158
17	50,817	0.14	181,531	179
18	43,113	0.20	108,518	177
19	12,615	0.80	36,583	229

Table 3 *Nucleating Co-citations.* For each of the 20 co-citation clusters, the pair with the strongest edge (greatest normalized co-citation frequency) is shown below along with a manually assigned label.

Cluster	Nucleating Pair	NCF	Manual Label
18670	10.1016/j.cam.2015.03.057 10.1016/j.sigpro.2015.10.009	0.92	Dynamical Systems
18892	10.1287/msom.1080.0228 10.1287/msom.1060.0190	0.72	Operations Research
18918	10.1007/s11128-010-0177-y 10.1007/s11128-013-0567-z	0.83	Image Processing
18947	10.1109/TASE.2011.2160452 10.1109/TASE.2011.2178023	0.77	Robotics
18956	10.1109/ICCV.2017.32 10.1109/ICCV.2017.31	0.84	Computer Vision
18972	10.1109/ASE.2013.6693094 10.1145/2568225.2568254	0.69	Software
18974	10.1016/j.amc.2009.03.023 10.1016/j.amc.2010.07.064	0.54	Applied Mathematics
18983	10.1137/110848864 10.1137/110848876	0.66	Optimization
18997	10.1504/IJMIC.2014.065339 10.1504/IJMIC.2015.068871	0.91	Chaotic Systems
19008	10.1016/j.chb.2008.12.013 10.1016/j.chb.2008.12.012	0.63	Hum Comp Interaction
19009	10.1109/AVSS.2017.8078491 10.1109/ICCV.2017.206	0.78	Artificial Intelligence
19013	10.1145/2508859.2516668 10.1007/978-3-642-42045-0_15	0.88	Security
19026	10.1145/2541940.2541942 10.1109/HPCA.2014.6835965	0.81	Architecture
19027	10.1016/S0305-0548(03)00250-8 10.1016/j.ejor.2006.03.013	0.82	Graph algorithms
19030	10.1109/ICDE.2008.4497474 10.14778/1687627.1687666	0.69	Databases
19031	10.1109/TMM.2005.843347 10.1109/TCE.2005.1405724	0.55	Networks
19033	10.1109/TIFS.2014.2327757 10.1109/TCYB.2014.2376934	0.78	Facial Recognition
19034	10.1016/j.neuroimage.2013.05.018 10.1016/j.neuroimage.2014.06.016	0.63	Neurology
19035	10.1002/int.21933 10.1002/int.21927	1.09	Intelligent Systems
19036	10.1006/jsco.2000.0402 10.1006/jsco.2000.0403	0.73	Comp Geometry

Table 4 *Nucleating Co-citations* Manually assigned labels for nucleating co-cited pairs (Table 3) are matched to the ASJC minor subject area that constitutes the largest fraction (shown as percentage in parentheses) of all nodes in the cluster. In cases of ties, both minor subject areas are shown. Abbreviations: Electrical and Electronic Engineering (EEE); Control and Systems Engineering (CSE).

Cluster	Label	Minor Subject Area
18670	Dynamical Systems	(i) EEE (14) (ii) CSE (14)
18892	Operations Research	EEE (24)
18918	Image Processing	(i) EEE (17) (ii) Software (17)
18947	Robotics	Software (12)
18956	Computer Vision	Computer Graphics and Computer-Aided Design (21)
18972	Software	Software (55)
18974	Applied Mathematics	Applied Mathematics (12)
18983	Optimization	EEE (35)
18997	Chaotic Systems	CSE (23)
19008	Hum Comp Interaction	Education (19)
19009	Artificial Intelligence	Computer Vision and Pattern Recognition (32)
19013	Security	Computer Science(all) (29)
19026	Architecture	EEE (35)
19027	Graph algorithms	Computer Networks and Communications (19)
19030	Databases	Software (19)
19031	Networks	EEE (32)
19033	Facial Recognition	Computer Vision and Pattern Recognition (18)
19034	Neurology	Neurology (12)
19035	Intelligent Systems	Artificial Intelligence (17)
19036	Comp Geometry	Software (10)

Acknowledgements The authors thank Henry Small for very helpful discussions. Research and development reported in this publication was partially supported by funds from the National Institute on Drug Abuse, National Institutes of Health, US Department of Health and Human Services, under Contract No HHSN271201800040C (N44DA-18-1216). TW is supported by the Grainger Foundation. Citation data used in this paper relied on Scopus (Elsevier Inc.) as implemented in the ERNIE project (Korobskiy et al., 2019), which is collaborative between NET ESolutions Corporation and Elsevier Inc. We thank our Elsevier colleagues for their support of the ERNIE project.

Conflict of interest

The authors declare that they have no conflicts of interest. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, NET ESolutions Corporation, or Elsevier Inc.

References

1. Almeida, H., Guedes, D., Meira Jr, W., Zaki, M.: Towards a better quality metric for graph cluster evaluation. Journal of Information and Data Management (JIDM) **3**, 378–393 (2012)
2. Association for Computing Machinery: Computing Classification System (2012). URL `https://dl.acm.org/ccs/ccs.cfm`. Accessed June 2019
3. Boyack, K., Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology **61**(12), 2389–2404 (2010). DOI 10.1002/asi.21419
4. Boyack, K.W.: Investigating the effect of global data on topic detection. Scientometrics **111**(2), 999–1015 (2017). DOI 10.1007/s11192-017-2297-y. URL `https://doi.org/10.1007/s11192-017-2297-y`
5. Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N., Börner, K.: Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. PLOS ONE **6**(3), e18029 (2011). DOI 10.1371/journal.pone.0018029. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018029`
6. Boyack, K.W., Patek, M., Ungar, L.H., Yoon, P., Klavans, R.: Classification of individual articles from all of science by research level. Journal of Informetrics **8**(1), 1–12 (2014). DOI 10.1016/j.joi.2013.10.005
7. Boyack, K.W., Small, H., Klavans, R.: Improving the accuracy of co-citation clustering using full text: Improving the Accuracy of Co-citation Clustering Using Full Text. Journal of the American Society for Information Science and Technology **64**(9), 1759–1767 (2013). DOI 10.1002/asi.22896. URL `http://doi.wiley.com/10.1002/asi.22896`
8. Chakraborty, T.: Role of interdisciplinarity in computer sciences: quantification, impact and life trajectory. Scientometrics **114**(3), 1011–1029 (2018). DOI 10.1007/s11192-017-2628-z. URL `https://doi.org/10.1007/s11192-017-2628-z`
9. Clarivate Analytics: Web of Science (2019). URL `https://clarivate.com/webofsciencegroup/solutions/web-of-science/`. Accessed Dec 2019
10. Dhillon, I., Guan, Y., Kulis, B.: Weighted graph cuts without eigenvectors: A multilevel approach. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 29:11, pp. 1944–1957. ACM Press (2007)
11. Elsevier: Scopus (2019). URL `https://www.scopus.com/home.uri`. Accessed Dec 2019
12. Emmons, S., Kobourov, S., Gallant, M., Börner, K.: Analysis of network clustering algorithms and cluster quality metrics at scale. PloS one **11**(7), e0159161 (2016)

13. Glänzel, W., Thijs, B.: Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. Scientometrics **111**(2), 1071–1087 (2017). DOI 10.1007/s11192-017-2301-6. URL https://doi.org/10.1007/s11192-017-2301-6
14. Kessler, M.M.: Comparison of the results of bibliographic coupling and analytic subject indexing. American Documentation **16**(3), 223–233 (1965). DOI 10.1002/asi.5090160309. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090160309
15. Klavans, R., Boyack, K.W.: Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? Journal of the Association for Information Science and Technology **68**(4), 984–998 (2017). DOI 10.1002/asi.23734
16. Korobskiy, D., Davey, A., Liu, S., Devarakonda, S., Chacko, G.: Enhanced Research Network Informatics Environment (ERNIE). Github repository, NET ESolutions Corporation (2019). URL https://github.com/NETESOLUTIONS/ERNIE
17. Marshakova-Shaikevich, I.: System of document connections based on references. Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy **6**(4), 3–8 (1973). DOI 10.1002/asi.4630240406
18. National Academies of Sciences, Engineering, and Medicine, et al.: Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments. The National Academies Press, Washington, DC (2018). DOI 10.17226/24926
19. National Science Foundation: Classification of Fields of Study (2012). URL https://www.nsf.gov/statistics/nsf13327/pdf/tabb1.pdf. Accessed June 2019
20. Perianes-Rodriguez, A., Ruiz-Castillo, J.: A comparison of the Web of Science and publication-level classification systems of science. Journal of Informetrics **11**, 32–45 (2017). DOI 10.1016/j.joi.2016.10.007
21. Salton, G., Bergmark, D.: A citation study of computer science literature. IEEE Transactions on Professional Communication **PC-22**(3), 146–158 (1979). DOI 10.1109/TPC.1979.6501740
22. Shu, F., Julien, C.A., Zhang, L., Qiu, J., Zhang, J., Larivière, V.: Comparing journal and paper level classifications of science. Journal of Informetrics **13**(1), 202–225 (2019). DOI 10.1016/j.joi.2018.12.005
23. Shun, J., Roosta-Khorasani, F., Fountoulakis, K., Mahoney, M.W.: Parallel Local Graph Clustering. Proc. VLDB Endow. **9**(12), 1041–1052 (2016). DOI 10.14778/2994509.2994522
24. Siebel, T.: Digital transformation: survive and thrive in an era of mass extinction. RosettaBooks (2019)
25. Sjögårde, P., Ahlgren, P.: Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. Quantitative Science Studies pp. 1–32 (2019). DOI {10.1162/qss_a_00004}. URL https://doi.org/10.1162/qss_a_00004
26. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science **24**(4), 265–269 (1973). DOI 10.1002/asi.4630240406
27. Small, H., Griffith, B.C.: The Structure of Scientific Literatures I: Identifying and Graphing Specialties. Science Studies **4**(1), 17–40 (1974). DOI 10.1177/030631277400400102
28. Small, H., Sweeney, E.: Clustering the science citation index using co-citations. Scientometrics **7**(3), 391–409 (1985). DOI 10.1007/BF02017157
29. The dblp Team: dblp Computer Science Bibliography (2018). URL https://dblp.org/xml/release/dblp-2018-08-01.xml.gz. Accessed June 2019
30. Traag, V.A., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports **9**(1), 1–12 (2019). DOI 10.1038/s41598-019-41695-z
31. Šubelj, L., van Eck, N.J., Waltman, L.: Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods. PLOS ONE **11**(4), e0154404 (2016). DOI 10.1371/journal.pone.0154404
32. Waltman, L., van Eck, N.J.: A new methodology for constructing a publication-level classification system of science. Journal of the American Society for Information Science and Technology **63**(12), 2378–2392 (2012). DOI 10.1002/asi.22748

33. Wang, Q., Waltman, L.: Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. Journal of Informetrics **10**(2), 347–364 (2016). DOI 10.1016/j.joi.2016.02.003